LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  |  |  |
| 2 | REVIEW OF LITERATURE  **Numerical Techniques to Represent Text Data**   1. 1-Hot Encoding 2. Bigram Model 3. Bag of Words 4. Vector Semantics 5. Word2Vec Sample Sentence 6. Updation Criteria of Word2Vec Matrix   **Summarization**   1. Text Summarization Approaches 2. Types of Extractive Summarization 3. PageRank Nodes and Edges 4. Singular Value Decomposition   **Evaluation of text summarization**   1. Types of Evaluations of Summaries 2. Algorithms for Intrinsic Evaluation 3. Types of ROUGE Measures   **Evaluation of text summarization**   1. Techniques for Web Scraping   **Twitter**   1. Trends as shown on Twitter on 27-03-2020 | 7  7  8  9  10  10  12  13  16  17  19  20  22  24  28 |
|  |  |  |
| 3 | ANALYSIS AND DESIGN   1. Component Diagram 2. Swimlane Diagram | 32  36 |
|  |  |  |
| 4 | IMPLEMENTATION DETAILS   1. Extractive Text Summarization 2. TextRank Algorithm 3. Latent Semantic Analysis Algorithm | 39  41  43 |
|  |  |  |
| 5 | RESULTS AND DISCUSSION   1. Average F-Measure Scores 2. F-Measure Scores Comparison 3. Average Computation Time 4. Bot Result on System Console on 13-03-2020 5. Bot Result on Twitter on 13-03-2020 6. Bot Result on Twitter on 27-03-2020 | 49  50  51  52  53  54 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  |  |  |
| 5 | RESULTS AND DISCUSSION   1. Zero Value Count in each Method 2. High F-Measure Probability per Method | 49  51 |

**LIST OF EQUATIONS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  |  |  |
| 2 | REVIEW OF LITERATURE   1. Word Frequency 2. Term Frequency 3. Inverse Document Frequency 4. TF-IDF 5. Recall 6. Precision 7. Brevity Penalty 8. BLEU Score | 14  14  15  15  19  20  21  22 |
|  |  |  |
| 3 | ANALYSIS AND DESIGN   1. F-Measure | 31 |
|  |  |  |
| 4 | IMPLEMENTATION   1. Improved Word Frequency | 42 |

**ABBREVIATIONS**

|  |  |
| --- | --- |
| **ABBREVIATION** | **DESCRIPTION** |
|  |  |
| TF-IDF | Term Frequency – Inverse Document Frequency |
| BLEU | Bilingual Evaluation Understudy |
| BP | Brevity Penalty |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |
| LCS | Longest Common Subsequence |
| WOEID | Where On Earth IDentifier |
| NLP | Natural Language Processing |
| NLTK | Natural Language Toolkit (Python) |
| API | Application Programming Interface |
| DOM | Document Object Model |
| GPU | Graphical Processing Unit |
| CPU | Central Processing Unit |
| RAM | Random Access Memory |
| XML | Extensible Mark-up Language |
| UML | Unified Modelling Language |
| JSON | JavaScript Object Notation |
| BBC | British Broadcasting Channel |
| URL | Uniform Resource Locator |
| TOI | Times of India |

**ABSTRACT**

With the boom of social media in the 21st century, it has grown to impact nearly every aspect of our life. This widespread nature of social media makes it a critical platform for the spread of information. This information, more often than not, is filled with flaws such as incomplete information, uncredited sources and may even go as far as being downright false. Platforms such as Facebook, Twitter, Instagram, and many more receive more than a billion active users every day with double that active every month, making them excellent platforms for the spread of information. There have always been suspicions and even some verified examples of these platforms being used to influence the lives of people by feeding them news from biased sources in order to influence the decisions they make in real life.[14]

Twitter is a critical player in the propagation of information with 500 million tweets containing snippets of information being sent every day, proved to be the ideal platform on which we could implement a system to verify and then spread information. The decision was made to consider the trending topics on twitter to be the keywords on which more information must be found and propagated.

Summarization of the selected articles, however, proved to be the greatest challenge. We decided to focus on extractive methods of summarization where sentences are lifted from the source verbatim instead of abstractive methods, which generated sentences due to its high complexity and resource requirements. We compared a variety of methods such as word frequency, term frequency-inverse document frequency, TextRank, latent semantic analysis based on the intrinsic evaluation. We also modified word frequency to obtain better results.

Bots proved to be the ideal form with which could implement said system as it reduces the probability of human error, removes work redundancy, and is inherently unbiased. The bot would scrape data off articles from reliable sources, summarize them using the best method from those mentioned above, and then tweet them along with a link to the actual source.